2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 978-1-6654-4899-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/CVPRW53098.2021.00113

Expression Transfer Using Flow-based Generative Models

Andrea Valenzuela
Universitat Pompeu Fabra
Barcelona, Catalonia / Spain
aand.valenzuela@gmail.com

Carlos Segura
Telefónica Research
Barcelona, Catalonia / Spain
carlos.seguraperales@telefonica.com

Ferran Diego
Telefónica Research, Spain
Barcelona, Catalonia / Spain
ferran.diegoandilla@telefonica.com

Vicenç Gómez
Universitat Pompeu Fabra
Barcelona, Catalonia / Spain
vicen.gomez@upf.edu

Abstract

Among the different deepfake generation techniques, flow-based methods appear as natural candidates. Due to the property of invertibility, flow-based methods eliminate the necessity of person-specific training and are able to reconstruct any input image almost perfectly to human perception. We present a method for deepfake generation based on facial expression transfer using flow-based generative models. Our approach relies on simple latent vector operations akin to the ones used for attribute manipulation, but for transferring expressions between identity source-target pairs. We show the feasibility of this approach using a pre-trained Glow model and small sets of source and target images, not necessarily considered during prior training. We also provide an evaluation pipeline of the generated images in terms of similarities between identities and Action Units encoding the expression to be transferred. Our results show that an efficient expression transfer is feasible by using the proposed approach setting up a first precedent in deepfake content creation, and its evaluation, independently of the training identities.

1. Introduction

The use of deep learning techniques to create fake multimedia is reshaping the field of multimedia forensics research [38, 26, 13, 16, 12]. Deepfake methods can easily create fake images and videos that cannot be distinguished from the authentic ones at the eyes of a human being [28]. Identity swap is the most popular application of these techniques, which consists of replacing the face image of a person, the target, with the face of another person, the source, to create an image or video of the target doing, and even saying, the things that the source does or says [29]. But identity swap is just one of many possible applications, including face synthesis, facial attribute manipulation, and expression transfer [28].

Autoencoders (AE) and generative adversarial networks (GAN) are the most popular architectures used for deepfake generation. These generative models excel in many tasks related to image synthesis such as face aging [3], attribute-guided face generation [24], or feature interpolation [37]. Despite their success, they typically require intensive person-specific training. Recently, flow-based deep generative models, also known as normalizing flows (NFs), have been proposed as an efficient alternative to the previous models [21]. NFs build on invertible transformations that map directly an input image into a latent representation that can be useful for high-quality imagery generation, facial attribute manipulation, and identity combinations by linear interpolation in the latent space.

In this work, we consider the flow-based *Glow* model [20] and explore how the original pre-trained model can be used for expression transfer between a source face image and a target face image. We make use of vector arithmetic operations in the latent space, and show that efficient expression transfer is feasible for source and target images corresponding to characters not included in the training set. We also propose an evaluation pipeline of the generated images in terms of how similar they are to the source and target identities and the quality of the facial expression itself based on Action Units.[1]

2. Deep Learning for Fake Multimedia

The first techniques for deepfake generation required significant manual editing until the first automatic method

[1] The results and implementation details for reproducing our experiments can be found at https://github.com/aandvalenzuela/normalizing-flows

978-1-6654-4899-4/21/$31.00 ©2021 IEEE
DOI 10.1109/CVPRW53098.2021.00113

was proposed [22]. Initially, 3D-based methods were used for transferring expressions from the source to the target face. In general, the transfer was performed by fitting a 3D morphable face model (3DMM) to both faces and then applying the expression components [30]. These type of statistical models were replaced by the so-called deep generative models, as deep learning entered the field of face synthesis and manipulation. Within this framework, most of the deepfakes are created using variations of autoencoders (AE) and generative adversarial networks (GAN), or a combination of both [28].

Recently, normalizing flows (NFs) have been shown to provide very successful results in several application domains [33, 31]. These models enjoy several properties that make them attractive for being used to generate deepfakes. In contrast to AEs and GANs, NFs are fully invertible and they allow for exact latent-variable inference and log-likelihood evaluation, while eliminating the need of subject-specific training (person- specific or pair-specific training) [2].

One of the first works proposing non-linear invertible transformations of the data to model complex high-dimensional densities for image synthesis was the non-linear independent component estimation (NICE) model [9] and real-valued non-volume preserving transformations (RealNVP) [10], an extension of NICE with a more flexible invertible transformation to experiment with natural images.

However, these flow-based generative models tended to perform worse in terms of density estimation compared to other generative models, and are incapable of realistic synthesis of large images compared to GANs. The *Glow* model [20], a generative flow with invertible 1×1 convolutions, significantly outperformed previous flow-based methods, both in density estimation on natural images, as well as in the ability to generate realistic high-resolution natural images efficiently. The *Glow* architecture has become the most popular flow-based generative model and has been subsequently modified with the objective of closing the performance gap between flow-based models and autoregressive models. Among these modifications, the most important ones are *Flow++* [15] and *MaCow* [25].

One of the main benefits of these techniques is the expressive power of their learned latent representations, since they can be used in downstream machine learning tasks, allowing for various applications such as interpolation between attributes and meaningful modifications of existing images. In [32], a stable model configuration was introduced for training deep Convolutional Neural Network (CNN) models as part of the GAN architecture. The authors explored the latent space of GANs with different training datasets, including a face database. They used the learned representations of the model to perform vector arithmetic with faces in the latent space.

Latent vector arithmetic has its origins in word-embedding models for natural language processing [27]. In word-embeddings, each word is assigned to a high-dimensional vector such that the geometry of the vectors captures semantic relations between the words, e.g. vectors being closer together has been shown to correspond to more similar words [17]. The most popular example is

$$h(\text{king}) - h(\text{man}) + h(\text{woman}) = h(\text{queen}),$$

where $h(w)$ is the latent vector corresponding to the word w. Analogously, an example proposed for vector arithmetic with faces is [32]: *smiling woman* $-$ *neutral woman* $+$ *neutral man* $=$ *smiling man*.

Specifically, the arithmetic was performed by averaging the points in the latent space of multiple faces with a given attribute. Therefore, the terms *smiling woman*, *neutral woman* and *neutral man* correspond to the resulting face after averaging multiple faces with the same attributes (*smile* and *neutral*). This approach introduced a methodology for transferring attributes, such as the smile, the color of the hair, complements as sunglasses, etc, into face images. This technique was later known as *attribute manipulation*.

The attribute manipulation approach can also be formulated as finding the path in the latent space present between the two extremes of the same attribute. In practice, the dataset is split into two subsets: the images with this concrete attribute (positive class samples) and the images without the attribute (negative class samples) to compute their corresponding average latent vectors, $\mathbf{z}_+$ and $\mathbf{z}_-$. Then, the resulting manipulation of an image $\mathbf{x}$ was obtained by adding to its latent vector $\mathbf{z}_{\text{input}}$ a scaled *manipulation vector*, consisting of the difference between the corresponding average encoding of each extreme. Therefore,

$$\mathbf{z}_{\text{manipulated}} = \mathbf{z}_{\text{input}} + \alpha \cdot \mathbf{z}_{\text{manipulation vector}}, \tag{1}$$

where $\mathbf{z}_{\text{manipulation vector}} = \mathbf{z}_+ - \mathbf{z}_-$ denotes a manipulation from the positive to the negative class. The coefficient α regulates the intensity of the manipulation, which typical ranges between -1 and $+1$. The sub-range $\alpha \in [-1, 0)$ is used for inverting the direction of the transformation, i.e. from the negative class to the positive class. For $\alpha = 0$, the original image is obtained.

Flow-based models have also explored this idea of attribute manipulation, e.g. for adding several attributes to their original images [20]. In this work, we want to continue with this semantic manipulation approach, but instead of using it for attribute manipulation, we explore whether same technique can be used to transfer expressions between a source identity and a target identity.

3. Normalizing Flows

A flow model f consists of an invertible transformation that maps the input image $\mathbf{x}$ to a standard Gaussian latent variable $\mathbf{z} = f(\mathbf{x})$ of the same dimensionality. The transformation f is created by stacking simple invertible transformations as $f(\mathbf{x}) = f_1 \circ \cdots \circ f_L(\mathbf{x})$, with each f_i having a tractable inverse and a tractable Jacobian determinant.

Optimizing the parameters of the flows f_i for a dataset of images involves maximizing the likelihood

$$\log p(\mathbf{x}) = \log \mathcal{N}(f(\mathbf{x}); \mathbf{0}, \mathbf{I}) + \sum_{i=1}^{L} \log \left| \det \frac{\partial f_i}{\partial f_{i-1}} \right|,$$

which can be done efficiently.

Once the model is trained, the process of generating a new image involves sampling from a Gaussian distribution and computing the inverse flow, i.e., $f^{-1}(\mathbf{z}) = f_L^{-1} \circ \cdots \circ f_1^{-1}(\mathbf{z})$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

In the Glow model, the individual flows f_i use an affine coupling layer and pixelwise reshuffling by a learned 1×1 convolution. For more details, see [20].

4. Expression Transfer

Our approach relies on simple vector arithmetic in the latent space for expression transfer between a source identity $\mathbf{x}_{src}$ and a target identity $\mathbf{x}_{tgt}$. The goal of the proposed vector arithmetic operation is to capture the facial expression of the source identity and synthesize a facial image $\mathbf{y}_{tgt}$ of the target identity with an analogous expression to the one of the source identity.

We start from a pre-learned Glow model and require additional (small) sets, $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$, of S images of the source face, and T images of the target face, respectively. We first compute the mean latent vector for both the source and target image sets, $\bar{\mathbf{z}}_{src} = \frac{1}{S} \sum_{\mathbf{x} \in \mathcal{D}_{src}} f(\mathbf{x})$ and $\bar{\mathbf{z}}_{tgt} = \frac{1}{T} \sum_{\mathbf{x} \in \mathcal{D}_{tgt}} f(\mathbf{x})$, respectively. Such vectors encode what we call the *mean* face of an identity, defined as its neutral and frontal (expressionless) face. We compute the output target image according to

$$\mathbf{y}_{tgt} = f^{-1}\left(\bar{\mathbf{z}}_{tgt} + \alpha \cdot (\mathbf{z}_{src} - \bar{\mathbf{z}}_{src})\right). \tag{2}$$

Similarly to equation (1) in the context of attribute manipulation, the vector $\mathbf{z}_{src} - \bar{\mathbf{z}}_{src}$ acts as an *expression vector*, and the coefficient α regulates the transfer intensity.

5. Experimental Results

For the experiments, we used the pre-trained Glow model on the *CelebA* dataset provided by *OpenAI*[2]. The *CelebA* dataset is a large-scale face attributes dataset with more than $200,000$ celebrity images from $10,177$ identities. The dataset contains useful annotations for manipulating the attributes of the synthesised faces [23]. The pre-trained Glow model has a dimensionality of $196,608$ and it is composed of 6 layers.

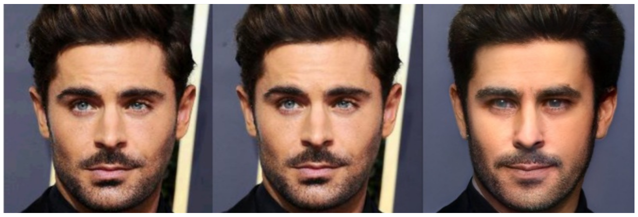

Figure 1. Example of face reconstruction using pre-trained models for a new face not observed during training. (left) Glow model, (middle) original image, and (right) StyleGAN2 model.

We consider five popular identities for illustrative purposes. Our choice considered both genders (two females and two males) and another identity with an special attribute (darker skin). For each subject, we crawled twenty additional face images from the Internet to form the small datasets $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$ ($S = T = 20$). The only processing for computing the mean latent vector is an eye alignment before applying the transformation f.

The proposed approach relies on the ability of the Glow model to reconstruct each of the new images not present in the training datasets, and to compute a valid expressionless face for each of the identities. Both tasks are the building blocks for the expression transfer approach.

5.1. Face Reconstruction

We first focus on the face reconstruction task, and compare the performance of the pre-trained Glow model with the StyleGAN2 network [19]. In this experiment, the StyleGAN2 network has been trained in a higher quality image dataset with an improved variability compared to the *CelebA* dataset.

Figure 1 shows an example of an original new image not present in the training datasets (middle), and the reconstructed images using the pre-trained Glow model (left) and the StyleGAN2 network (right). The Glow model encodes the new image into the latent space and decodes it back to the exact same image, leading to a perfect reconstruction, in agreement with [4]. In contrast, the StyleGAN2 reconstruction is less accurate. The perfect reconstruction is a direct consequence of the Glow model being invertible.

There are hybrid architectures combining an autoencoder with the adversarial architecture, but they still present difficulties in the reconstruction process. Although there exist several efficient embedding algorithms that can map a given image into the extended latent space of a pretrained StyleGAN2 network, they still present problems generalizing beyond the training dataset [1].

5.2. Mean of a Person

The second building block of our approach for expression transfer is the computation of a valid average face for

[2] https://github.com/openai/glow

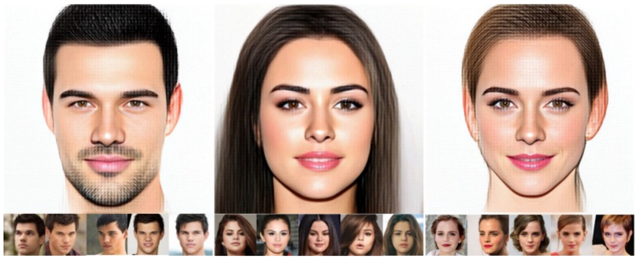

Figure 2. Examples of mean faces of three public identities. (top) The three main faces. (bottom) Samples from the set of 20 real images used to compute the mean.

both the source and the target. As previously mentioned, we refer to such expressionless image as the *mean* face of a person. We can directly obtain such an image by inverting the flow from the mean latent vector. For example, $\bar{\mathbf{x}}_{src} = f^{-1}(\bar{\mathbf{z}}_{src})$ in the case of the source image.

Figure 2 shows three illustrative examples of mean faces considering the original twenty crawled images of each identity. We observe that, despite the heterogeneity present in the datasets, the model successfully decodes a relatively neutral and frontal face without loosing the identity of the person under consideration. In some cases, the neutral face may present a residual expression, as the slight smile of the third identity in Figure 2, that could be explained by bias in the data.

5.3. Arithmetic Expression Transfer

We now focus on the task of expression transfer and analyze the approach proposed in equation (2) for the range of values of the coefficient α originally considered for the attribute manipulation task, $\alpha \in [0, 1]$.

Figure 3 shows an example. The first image (top-left) corresponds to the source input image $\mathbf{x}_{src}$ and contains the expression to be transferred. The rest of the images are the resulting $\mathbf{y}_{tgt}$ for increasing values of α. Thus, the second image (top row) corresponds to $\alpha = 0$ (target mean image $\bar{\mathbf{x}}_{tgt}$) and the final image (bottom-right) corresponds to $\alpha = 1$. The original exploration was performed in α-steps of 0.1 between consecutive samples.

We observe a trade-off between quality of the obtained image and the associated facial expression. Lower values of α result in images too close to the mean target face, whereas higher values of α result in poor quality faces. For values of $\alpha \in [0.3, 0.5]$, we identify an efficient expression transfer.

Figure 4 shows further results for a different character and five different expressions (rows), varying α in the aforementioned narrower range in steps of 0.02 between consecutive samples. Again, every first image of each row corresponds to the original image with the expression to be transferred, $\mathbf{x}_{src}$. The second image of each row corresponds to the mean face of the target identity $\bar{\mathbf{x}}_{tgt}$, while the other images correspond to the target output images $\mathbf{y}_{tgt}$ for increasing values of α.

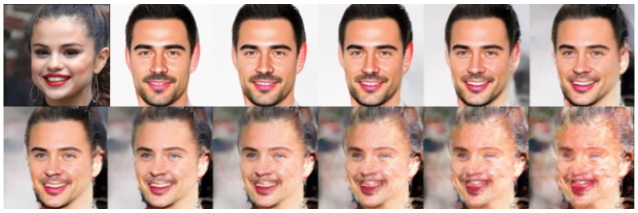

Figure 3. Expression transferring between two identities. The first image of the first row corresponds to the expression to be transferred, while the other images correspond to the target output image generated for increasing values of alpha within the range $[0, 1]$ (the second image of the first row corresponds to the target mean image).

The proposed approach is also valid for transferring the same facial expression to different target identities. To illustrate this multi-target ability, Figure 5 shows an example of the same expression transfer applied to different identities. It shows three different target identities (rows) where the first image of each row corresponds to the original source input image $\mathbf{x}_{src}$ with the expression to be transferred. The second image of each row corresponds to the target mean image $\bar{\mathbf{x}}_{tgt}$ of the identities under consideration, while the other images correspond to the target output images $\mathbf{y}_{tgt}$ for increasing values of α within the range $\alpha \in [0.3, 0.5]$ again in steps of 0.02 between consecutive samples. In these illustrative examples, almost all the inner faces of the intermediate samples obtained within the proposed range of α look like realistic faces.

From these results, we can conclude that expression transfer using a pre-trained Glow model is feasible using our proposed approach.

6. Optimizing α for Expression Transfer

So far, we have presented results that show the feasibility of the proposed approach on a small set of illustrative identities. In this section, we describe two alternative methods that aim to find the best expression transfer using our proposed approach regardless of the identity. These methods optimize two different metrics as a function of α, and can additionally be useful for evaluation purposes.

6.1. Identity Similarity

Our first method relies on a measure of identity similarity between two images. We compare how much similar is the generated image generated $\mathbf{y}_{tgt}$ with two other images: the mean face of the target identity $\bar{\mathbf{x}}_{tgt}$ (target distance) and the source identity $\mathbf{x}_{src}$ (source distance). Intuitively, these similarities should change consistently as a function of α across identities.

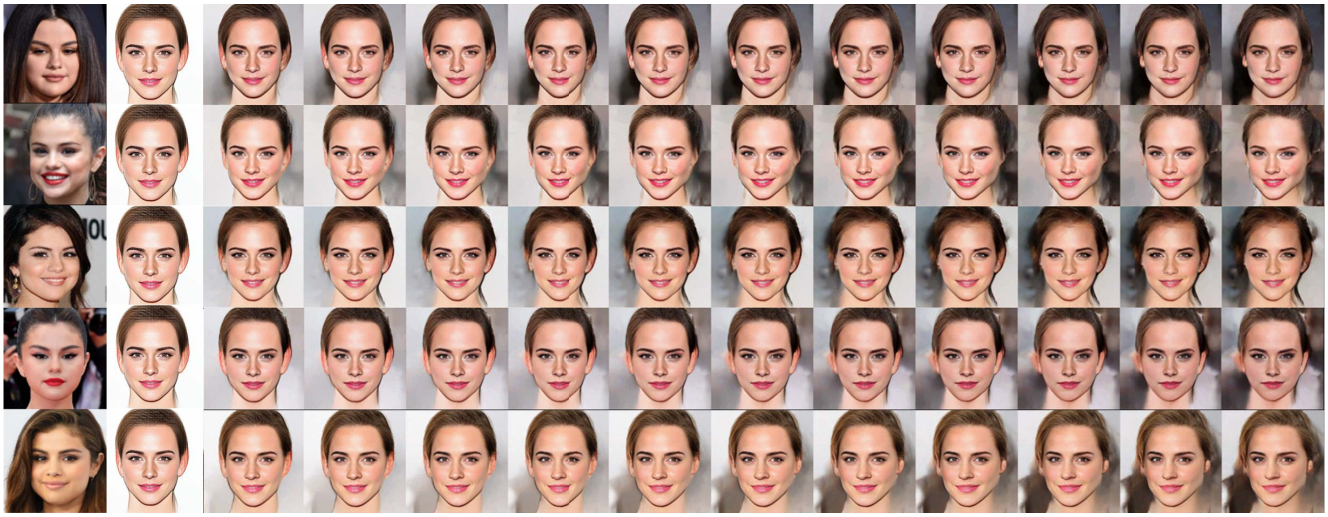

Figure 4. Expression transfer: five different expressions for the same target identity. The first image of each row corresponds to the original expression to be transferred. The second image of each row corresponds to the mean face of the target identity, while the images next to them show the target output images for increasing values of α. In this case, the range $\alpha \in [0.3, 0.5]$ is considered with steps of 0.02 between consecutive samples.

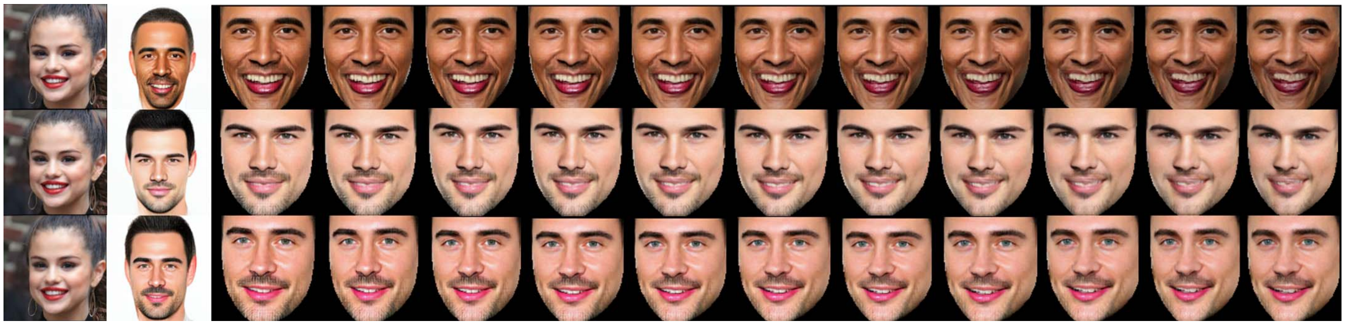

Figure 5. Expression transfer: same expression for three different target identities. The first image of each row corresponds to the original expression to be transferred. The second image of each row corresponds to the mean face of the target identities, while the images next to them show the target output images for increasing values of α. In this case, the range $\alpha \in [0.3, 0.5]$ is considered with steps of 0.02 between consecutive samples.

To compute both similarities, we use the CNN Inception Resnet (V1), that combines the residual connections introduced in [14] and the latest revised version of the Inception architecture [35]. In particular, we evaluate the target output images using the implementation for Inception Resnet (V1) models in `pytorch`[3] pretrained on both `VGGFace2` [7] and `CASIA-Webface` datasets [39]. This implementation of the Inception Resnet returns a distance score for each pair of input images. Lower values for the distance mean that the identities of the images look alike, and vice-versa.

Figure 6 shows the distances corresponding to four different identities for $\alpha \in [0, 1]$. We observe that the source distance (discontinuous line) remains generally constant, meaning that the expression vector successfully gets rid of the source identity components. In contrast, the target distance (continuous line) increases with α. [4] This is a general trend, and we typically observe that beyond $\alpha \approx 0.6$, the generated image deviates more from the target mean than from the source identity, suggesting a value of α smaller than this threshold to preserve the identity of the target in the expression transfer. Note that this is in agreement with the previously identified expression transfer examples, for which we identified a range of $\alpha \in [0.3, 0.5]$.

[3] https://github.com/timesler/facenet-pytorch

[4] We checked distances between several real images of the same identity and found values around 0.8.

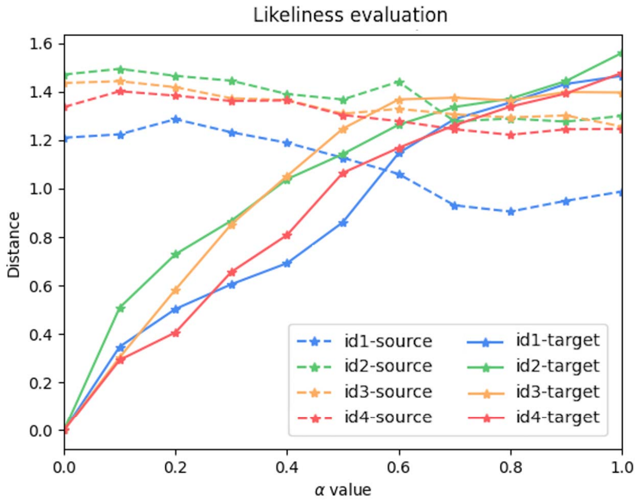

Figure 6. Evolution of the distance score of the generated samples with respect to the source mean (discontinuous line) and the target mean (continuous line). Despite the fact that the distance score with respect to the source remains almost constant for any value of α, the distance score with respect to the target mean increases for increasing values of α.

6.2. Expression Characterization Using Action Units

Our second method relies in the particular expression to be transferred. For that, we characterize the expression quantitatively and compute the deviation between such a characterization for the source identity $\mathbf{x}_{src}$ and the generated image $\mathbf{y}_{tgt}$.

The Facial Action Coding System (FACS) [11] is a comprehensive anatomically-based system for describing all visually discernible facial movement. FACS was created with the purpose of identifying the different facial muscle movements that correspond to a certain emotion to objectively determine the displayed emotion of a given face. Such analysis of facial expressions is one of very few techniques available for assessing emotions in real-time [8].

This method breaks down facial expressions into individual components of muscle movement. Each one of these individual components is called Action Unit (AU) and a combination of AUs represents a particular emotion [36]. For example, happiness is calculated from the combination of AU6 (cheek raiser) and AU12 (lip corner puller).

To evaluate the expression itself of the target output images $\mathbf{y}_{tgt}$, we use the toolkit OpenFace 2.0. This toolkit is freely available and reaches state-of-the-art performance in all of the above mentioned tasks [5]. OpenFace 2.0 recognizes facial expressions by detecting both the intensity and the presence of certain AUs. The intensity is on a 5 point scale (0: not present to 5: maximum intensity) representing the degree of activation [6]. Using this toolkit it is possible to detect the eighteen AUs listed in Table 1.

Description of the Action Units			
AU	**Full name**	**AU**	**Full name**
1	Inner brow raiser	14	Dimpler
2	Outer brow raiser	15	Lip corner depressor
4	Brow lowerer	17	Chin raiser
5	Upper lip raiser	20	Lip stretched
6	Check raiser	23	Lip tightener
7	Lid tightener	25	Lips part
9	Nose wrinkler	26	Jaw drop
10	Upper lip raiser	28	Lip suck
12	Lip corner puller	45	Blink

Table 1. Action Units analyzed using OpenFace 2.0. Table shows the eighteen AUs used to characterize the transferred expressions.

We propose the following approach to evaluate the generated faces: the facial expression of the original image of the source $\mathbf{x}_{src}$ is characterized using the eighteen AUs and their intensity scores are treated as the ground truth (GT) for the expression to be transferred. A generated image $\mathbf{y}_{tgt}$ is then evaluated as well by characterizing the expression in terms of the intensity of the eighteen AUs and comparing these values to the GT.

As an example, Figure 7 shows the characterization of one of the considered expressions in terms of presence of AUs and Figure 8 shows the analysis of this concrete expression transfer for the target identity of Figure 4. For each AU, we show the intensity values for both the minimal ($\alpha = 0.02$) and maximal ($\alpha = 1.00$) contributions of the expression vector. We also show the intensity of each AU for the value of α minimizing the error with respect to the GT ($\alpha^* = 0.44$). These three representations (discontinuous lines) are presented in comparison to the GT (continuous line). It is interesting to note how the intensity profile corresponding to α^* gets closer to the GT presenting almost the same AU activation with relatively small intensity differences. In general, we do not observe a clear tendency showing an optimal value of α consistent across the considered identities and expressions. Nevertheless, the general trend shows that the values of α^*, namely the value of α minimizing the error with respect to the GT, are generally within the $\alpha \in [0.3, 0.5]$ range in which we visually identified the expression transfer.

The proposed AU framework can be used as an automatic procedure to find the target output image $\mathbf{y}_{tgt}$ that minimizes the error with respect to the GT. Across the different expressions to be transferred, we realized that there are certain AUs that present higher error rates in all the transfers, suggesting that this concrete aspect of the expression is difficult to either be captured in the expression vector, $\mathbf{z}_{src} - \bar{\mathbf{z}}_{src}$, or to be modified in the source mean, $\bar{\mathbf{x}}_{src}$. This is the case, e.g., of AU45, which captures the level of

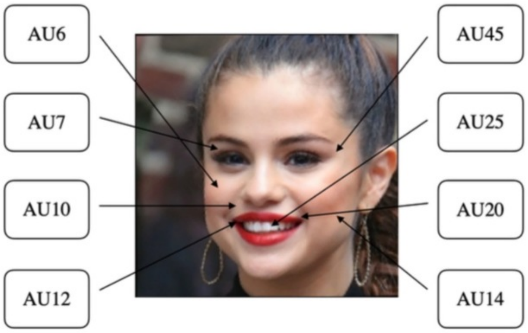

Figure 7. Example of an original expression to be transferred, $\mathbf{x}_{src}$, analysed in terms of Action Units. The AUs present in the expression are marked regardless of their intensity.

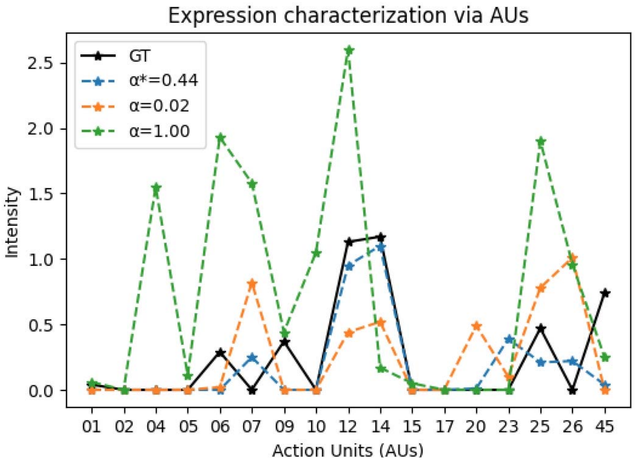

Figure 8. AU's intensity analysis for a concrete expression transfer. The original expression (black line) shows the GT intensity pattern. The extreme values of the α coefficient have been represented (orange line for $\alpha = 0.02$ and green line for $\alpha = 1.00$) as well as the value of α minimizing the error, α^* (blue line).

blinking in the facial expression. The nose wrinkle ($AU9$) and the jaw drop ($AU26$) also present the same problem.

Regarding the target identities, we have seen that the error increases considerably when target identities have a racial attribute different form the one classified as *white race*. According to [18], we hypothesize that these phenomenon could be present because of the strong bias that the *CelebA* dataset exhibits towards Caucasian faces, i.e., racially classified as *white*. The racial composition of the *CelebA* dataset is of 80% white race, while the other 20% is shared between the other races (the *black* race represents less than the overall 10%) [18, 34].

7. Discussion and Conclusion

We have proposed a simple vector arithmetic approach for expression transfer between two concrete identities using normalizing flows. The proposed approach relies on the ability of the model to encode and decode original images regardless of their presence or not in the training dataset, and in the ability to generate expressionless face images without loosing the identity of the person under consideration. The pre-trained Glow model has been validated according to these two properties due to the fact that it is a flow-based model.

Our proposed formula is analogue to the one traditionally used for attribute manipulation, where the original *manipulation vector* is redefined as an *expression vector*, computed as the difference between the source input and the source mean (instead of the average latent vectors of the two extremes of the attribute to modify). In addition, instead of modifying an original image, we propose the application of the expression vector to the expressionless image of the target identity.

Our initial study considers five different identities and a limited number of expressions, but already provides valuable results for future large-scale analysis and validation. In particular, we have found a valid range of values for the parameter $\alpha \in [0.3, 0.5]$ that controls the expression transfer.

The generated images of the target identity have been analysed in terms of likeliness to the target and source identities via the comparison to the mean faces. We have shown that in all the cases the identity of the target is preserved during the expression transfer within the identified range of the parameter α.

In addition, the generated images have been also analysed in terms of the quality of the expression transfer via Action Units. Although there is not a clear tendency in the value of α showing a smaller error consistently between identities and expressions, this evaluation method also shows that the value of α^* generally lies within the identified range as well.

Overall, we have presented a successful expression transfer method for flow-based models that, because of their invertibility property, could be extrapolated to a variety of target identities regarding different facial expressions of the source aside from the training process. This work represents a first insight in deepfake creation via expression transfer in flow-based models providing an initial pipeline for both generation and evaluation of such type of deepfake content.

Acknowledgements

V. Gómez has received funding from "La Caixa" Foundation (100010434), agreement LCF/PR/PR16/51110009 and is supported by the Ramon y Cajal program RYC-2015-18878 (AEI/MINEICO/FSE,UE).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN

latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 3

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 38–45, 2019. 2

[3] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093, 2017. 1

[4] Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 399–409, 2020. 3

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 59–66, 2018. 6

[6] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6, 2015. 6

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 67–74, 2018. 5

[8] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007. 6

[9] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, Workshop Track Proceedings*, 2015. 2

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations*, 2017. 2

[11] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976. 6

[12] Fausto Galvan. Image/video forensics. *European Law Enforcement Research Bulletin*, (20):105–123, 2020. 1

[13] Luca Guarnera, Oliver Giudice, Cristina Nastasi, and Sebastiano Battiato. Preliminary forensics analysis of deepfake images. In *International Annual Conference (AEIT)*, pages 1–6, 2020. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2722–2730, 2019. 2

[16] Mousa Tayseer Jafar, Mohammad Ababneh, Mohammad Al-Zoube, and Ammar Elhassan. Forensics and analysis of deepfake videos. In *11th International Conference on Information and Communication Systems (ICICS)*, pages 053–058, 2020. 1

[17] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, chapter 19, pages ISBN 978–0–13–187321–6. 2009. 2

[18] Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558, 2021. 7

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 3

[20] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, pages 10215–10224, 2018. 1, 2, 3

[21] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1

[22] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015. 3

[24] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[25] Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard Hovy. Macow: Masked convolutional generative flow. In *Advances in Neural Information Processing Systems*, pages 5893–5902, 2019. 2

[26] Owen Mayer and Matthew C. Stamm. Exposing fake images with forensic similarity graphs. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1049–1064, 2020. 1

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, 2013. 2

[28] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 2021. 1, 2

[29] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection: A survey. *arXiv preprint arXiv:1909.11573*, 2019. 1

[30] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE international conference on computer vision*, pages 7184–7193, 2019. 2

[31] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, 2019. 2

[32] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR*, 2016. 2

[33] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1530–1538, 2015. 2

[34] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017. 7

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[36] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001. 6

[37] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017. 1

[38] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 1

[39] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5